Received: 20 March 2018 | Accepted: 20 September 2018

DOI: 10.1111/2041-210X.13103

RESEARCH ARTICLE

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

Automatic acoustic detection of birds through deep learning: The first Bird Audio Detection challenge

Dan Stowell[1] | Michael D. Wood[2] | Hanna Pamuła[3] | Yannis Stylianou[4] | Hervé Glotin[5]

[1]Machine Listening Laboratory, Centre for Digital Music, Queen Mary University of London, London, UK

[2]Ecosystems and Environment Research Centre, School of Environment and Life Sciences, University of Salford, Salford, UK

[3]Department of Mechanics and Vibroacoustics, AGH University of Science and Technology, Krakow, Poland

[4]Computer Science Department, University of Crete, Crete, Greece

[5]University Toulon, Aix Marseille University, CNRS, LIS, DYNI Team, SABIOD, Marseille, France

Correspondence
Dan Stowell
Email: dan.stowell@qmul.ac.uk

Funding information
Akademia Górniczo-Hutnicza im. Stanislawa Staszica, Grant/Award Number: 15.11.130.642; Engineering and Physical Sciences Research Council, Grant/Award Number: EP/L020505/1; Natural Environment Research Council, Grant/Award Number: NE/L000520/1; Radioactive Waste Management Ltd; ERASMUS+

Handling Editor: David Orme

Abstract

1. Assessing the presence and abundance of birds is important for monitoring specific species as well as overall ecosystem health. Many birds are most readily detected by their sounds, and thus, passive acoustic monitoring is highly appropriate. Yet acoustic monitoring is often held back by practical limitations such as the need for manual configuration, reliance on example sound libraries, low accuracy, low robustness, and limited ability to generalise to novel acoustic conditions.
2. Here, we report outcomes from a collaborative data challenge. We present new acoustic monitoring datasets, summarise the machine learning techniques proposed by challenge teams, conduct detailed performance evaluation, and discuss how such approaches to detection can be integrated into remote monitoring projects.
3. Multiple methods were able to attain performance of around 88% area under the receiver operating characteristic (ROC) curve (AUC), much higher performance than previous general-purpose methods.
4. With modern machine learning, including deep learning, general-purpose acoustic bird detection can achieve very high retrieval rates in remote monitoring data, with no manual recalibration, and no pretraining of the detector for the target species or the acoustic conditions in the target environment.

KEYWORDS
bird, deep learning, machine learning, passive acoustic monitoring, sound

1 | INTRODUCTION

World-wide, bird populations have exhibited steep declines since the 1970s, largely due to changes in land management (North American Bird Conservation Initiative, 2016; RSPB, 2013). Bird populations are also expected to change in number and distribution as the impacts of climate change play out in coming years (Johnston et al., 2013). It is thus crucial to monitor avian populations for the purposes of conservation, scientific research, and ecosystem management. This has traditionally been performed via manual surveying, often including the use of volunteers to help address the challenges of scale (Johnston et al., 2014; Kamp, Oppel, Heldbjerg, Nyegaard, & Donald, 2016). However, manual observation remains limited, especially in areas that are physically challenging to access, or when the focus is night-time behaviour. Many bird species are readily detectable by their sounds, often more so than by vision, and so with modern remote monitoring stations able to capture continuous audio recordings the prospect opens up of massive-scale spatio-temporal monitoring of birds (Aide et al., 2013; Frommolt, 2017; Furnas & Callas, 2015; Hill et al., 2017; Knight et al., 2017; Matsubayashi et al., 2017).

The first wave of such technology performed automatic recording but not automatic detection, relying on manual after-the-fact study of sound recordings (Frommolt, 2017; Furnas & Callas, 2015). Later projects have employed some form of automatic detection, which might be based on low-complexity signal processing such as energy thresholds or template matching (Colonna, Cristo, Júnior, & Nakamura, 2015; Towsey, Planitz, Nantes, Wimmer, & Roe, 2012), or on machine learning algorithms (Aide et al., 2013). However, when used for field deployments, practitioners face a common hurdle. With the current state of the art, all methods require manual tuning of algorithm parameters, customisation of template libraries, and/or post-processing of results, often necessitating some degree of expertise in the underlying method. The methods are not inherently able to generalise to new conditions—whether those conditions be differing species balances, noise conditions, or recording equipment. Many methods also exhibit only moderate accuracy, which is tolerable in small surveys but leads to unfeasible amounts of false negatives and -positives in large surveys (Marques et al., 2012). A further common limitation is the lack of robustness in particular to weather noise: sounds due to rain and wind are commonly observed to dramatically affect detector performance, and as a result, surveys may need to treat weather-affected recording periods as missing data.

Recent decades have witnessed extremely strong growth in the abilities of machine learning. The advances are not only due to increased dataset sizes and computational power but also due to deep learning methods that can learn to make predictions in extremely nonlinear problem settings, such as speech recognition or visual object recognition (LeCun, Bengio, & Hinton, 2015). These methods have indeed been applied to bioacoustic audio tasks (Goëau, Glotin, Vellinga, Planque, & Joly, 2016; Knight et al., 2017; Salamon & Bello, 2017; Salamon, Bello, Farnsworth, & Kelling, 2017), and it is clear that their use could enable many organisations to work more cheaply and efficiently (Joppa, 2017). However, even with the strong performance of modern machine learning, there remain important questions about generalisability (Knight et al., 2017). Machine learning workflows use a "training set" of data from which the algorithm "learns," optionally a "validation set" used to determine when the learning has achieved a satisfactory level, and then a "testing set" which is used for the actual evaluation to estimate the algorithm's typical performance on unseen data. Such evaluation is typically performed in matched conditions, meaning the training and testing sets are drawn from the same pool of data, and thus, general properties of the datasets—such as the number of positive vs. negative cases—are expected to be similar. This enables users to test that the algorithm can generalise to new items drawn from the same distribution. However, in practical deployments of machine learning, the new items are rarely drawn from the same distribution: conditions drift or the tool is applied to new data for which no training data are available (Knight et al., 2017; Sugiyama & Kawanabe, 2012). This is one reason that accuracy results obtained in research papers might not translate to the field.

In order to address such problems, we designed a public evaluation campaign focused on a highly general version of the bird detection task, intended specifically to encourage detection methods which are able to generalise well: agnostic to species and able to work in unseen acoustic environments. In this work, we present the new acoustic datasets which we collated and annotated, the design of the challenge, and its outcomes, with new deep learning methods able to achieve strong results despite the difficult task. We analyse the submitted system outputs for their detection ability as well as their robust calibration; we perform a detailed error analysis to inspect the sound types that remain difficult for machine learning detectors, and apply the leading system to a separate held-out dataset of night flight calls. We conclude by discussing the new state of the art represented by the deep learning methods that excelled in our challenge, the quality of their outputs, and the feasibility of deployment in remote monitoring projects.

2 | MATERIALS AND METHODS

To conduct the evaluation campaign, we designed a detection task to be solved—specific but illustrative of general-purpose detection issues—gathered multiple datasets and annotated them, and then led a public campaign evaluating the results submitted by various teams. After the campaign, we performed detailed analysis of the system outputs, inspecting questions of accuracy, generality, and calibration.

Our aim to facilitate general-purpose robust bird detection, agnostic to any specific application, was a key to how we designed the challenge specification. The task of "detecting" birds in audio can be operationalised in multiple ways: for example, a system that emits a trigger signal in continuous time representing the onset of each bird call, a system that identifies regions of pixels in a spectrogram representation (time–frequency "boxes"), or a system that estimates the number of calling individuals in a given time region (Benetos, Stowell, & Plumbley, 2018). For any given application, the choice of approach will depend on the requirements for downstream processing. We selected an option which we consider gave wide relevance, while also being a task that could be solved by diverse methods, from simple energy detection, through to template matching or machine learning. This was that audio should be divided into 10-s clips, and the task specification would be to label each clip with a binary label indicating the presence or absence of birds.

This approach quantises time such that any positive detection should be time localisable within ±10 s, which is sufficient for most purposes. It also restricts such that there is no indication of the absolute number of bird calls detected within a positively labelled clip; however, this is hard to ground-truth accurately. Also, via statistical ecology methods relative abundances may still be inferred from the distribution of positive detections (Marques et al., 2012). A concrete advantage of this approach was that it was much quicker to gather manual data annotations than would be the case for more complex labelling.

2.1 | Datasets

We gathered and annotated datasets from multiple sources. The purpose of this was twofold: first, to provide better evaluation of the generality of algorithms, and second, to provide challenge

TABLE 1 Recording locations in Chernobyl dataset

Codename	Habitat	Radiation
Buryakovka	Abandoned village	Low
S2	Deciduous forest	Medium
S11	Meadow area	Medium
S37	Pine forest	High
S60	Shrub area	Low
S93	Mixed forest	High

participants with development data (e.g., to perform trial runs or to train machine learning algorithms) in addition to testing data.

We used audio data from remote monitoring projects and also from crowdsourced audio recordings. These two dataset types differ from each other in many ways, such as: remote monitoring audio was passively gathered, while crowdsourced audio recordings were actively captured; the ratio of positive and negative items was different; remote monitoring used fixed and known recording equipment, while crowdsourcing used uncontrolled equipment. These differences were deliberately introduced for their use in ensuring that the challenge would be a strong test of generalisation.

Chernobyl dataset: Our primary remote monitoring dataset was collected in the Chernobyl Exclusion Zone (CEZ) for a project to investigate the long-term effects of the Chernobyl accident on local ecology (Gashchak, Gulyaichenko, Beresford, & Wood, 2017; Wood & Beresford, 2016). The project had captured over 10,000 hr of audio since June 2015, across various CEZ environments, using Wildlife Acoustics Song Meter 3 (SM3) units mounted at approximately 1.5 m above the ground. For the present work, we selected six recording locations representing different environments (Table 1), and from those selected a deterministic subsample: continuous 5-min audio segments at hourly intervals, across multiple days. Annotators manually labelled all time intervals in which birds were heard (using Raven Pro software), and then, we split recordings and metadata automatically into 10-s segments. The number of files per location is uneven because of limited annotator time, giving us 6,620 items in total. No weather filtering or other rejection of difficult regions was applied.

Warblr dataset: Our first crowdsourced dataset came from a UK-wide project Warblr. Warblr is a software application available for Android and Apple smartphones, which offers automatic bird species classification (using the method of Stowell and Plumbley (2014a)) for members of the public via the submission of 10-s audio recordings. We extracted a dataset of 10,000 audio files gathered in 2015–2016. The audio files were thus actively collected, recorded on diverse mobile phone devices, and likely to contain various human noises such as speech and handling noise. No assumptions can be made that the data were a representative sample of geographical locations, weather conditions, or bird species. Metadata for the files indicated that they covered all the UK seasons, many times of day (with a bias towards weekends and mornings) and geographically spread all around the UK, with a bias towards population centres.

All recordings were selected that fell within the time window of available data, limited to a maximum of 10,000. No selection or filtering of the data were performed beyond the self-selection inherent in crowdsourcing.

Although the data included automatic estimates of which bird species were present, these were not precise enough to be converted to ground-truth data for the detection challenge. We thus performed manual annotation, with each item being labelled as positive or negative according to the challenge specification. Most items were single annotated, although we were able to obtain double annotation for a small number of items, which allowed us to estimate inter-rater reliability. Annotation was performed by experienced listeners using headphone listening and a simple web interface.

freefield1010 dataset: Our second crowdsourced dataset was an existing public dataset called freefield1010 (Stowell and Plumbley, 2014b). This consists of 7,690 audio clips selected from the Freesound online audio archive. To create this dataset, the audio clips had been selected such that they were labelled with the "field-recording" tag in the database, and trimmed to 10 s duration. The data were of different origins than Warblr: they covered a global geographical range, and the recording devices used were almost never documented, but likely to include handheld audio recorders as used by pro-amateur sound recordists, as well as some mobile phones and some higher end recording devices. The Freesound database is crowdsourced and thus largely uncontrolled.

These data did not come with labels suitable for our challenge; instead, each item came with a set of freely chosen tags to indicate the content generally. We investigated the "birdsong" tag, one of the most commonly used (2.6% of items), but found this insufficiently accurate. We therefore had these audio annotated through the same process as the Warblr data.

PolandNFC dataset: The last dataset contains recordings from one author's (HP) project of monitoring autumn nocturnal bird migration (Pamuła, Klaczynski, Remisiewicz, Wszolek, & Stowell, 2017). The recordings were collected every night, from September to November 2016 on the Baltic Sea coast, near Darlowo, Poland. We used Song Meter SM2 units with weather-resistant, directional Night Flight Calls microphones from Wildlife Acoustics Inc., mounted on 3- to 5-m poles. The amount of collected data (>3,200 hr of recordings) exceeded what a human expert can annotate manually in reasonable time. Therefore, we subjectively chose and manually annotated a subset consisting of 22 half-hour recordings from 15 nights with different weather conditions and background noise including wind, rain, sea noise, insect calls, human voice, and deer calls. No other selection criterion or weather filtering was applied. Manual annotation was performed by visual inspection of a spectrogram and listening to the audio files. Only the passerine migrant calls were annotated (voices in 5–10 kHz range), so it may happen that some low-pitched bird species (e.g., resident owl calls) obtained no-bird label. However, such calls were extremely rare in the described dataset, so this could not have a strong effect on results.

Nocturnal bird calls are typically 10–300 ms duration, so a clip length different from other datasets was chosen; the selected recordings were split into 1-s clips. More details about the dataset structure (and analysis of the effect of audio clip duration) may be found in Pamuła et al. (2017).

All sound files used in the public challenge were normalised in amplitude and saved as a 16-bit single-channel WAV files at 44.1-kHz sampling rate (Normalisation via the sox tool using gain *−n* −2), and are available under open licences (see Data Accessibility statement).

Our data annotation process was designed after early community discussions about how the challenge should be conducted. We resolved that the annotations should reflect plausible annotation conditions as encountered in applications. In particular, they should be well-annotated; yet, any mislabellings discovered in the ground-truth data as the challenge progressed should not be eliminated, since training data in practice do contain some errors and are not subject to the same scrutiny as in a data challenge. A good detection algorithm must be able to cope with a small level of imprecision in the annotation data.

However, it was possible at the end of the challenge to perform further analysis and inspect the degree of machine errors and human errors. To make good use of annotator time, we used mismatch between automatically inferred decisions and manual annotations to search for mislabelled items in the dataset. For this, we used the mean decision from the strongest three submissions to the challenge. All items in the testing set with a negative ground-truth label but a mean decision >0.2 and all items with a positive ground-truth label but a mean decision <0.3 were examined and relabelled if needed. One might expect the threshold for revalidation to be 0.5: the asymmetry is because systems generally exhibited a bias towards low confidence, as will be seen later (Section 3.1). This revalidation process not only refined the testing set, but also allowed us to calculate a value for the inter-rater agreement for manual annotation, which we will express as an area under the receiver operating characteristic (ROC) curve (AUC) for comparison against the results of automatic detection. Note that the revalidation process requires the time of expert listeners, and so, it was not feasible to perform mass crowdsourcing on the whole collection.

2.2 | Baseline classifiers

To establish baseline performance against which to compare new methods, we used two existing machine learning based classification algorithms.

The first (code-named smacpy) was the same baseline classifier as used in a 2013 challenge on "detection and classification of acoustic scenes and events" ("DCASE") (Stowell, Giannoulis, Benetos, Lagrange, & Plumbley, 2015). This baseline classifier represents a well-studied method used in many audio classification tasks: audio is converted to a representation called mel frequency cepstral coefficients (MFCCs), and the distributions of MFCCs are then modelled using Gaussian mixture models (GMMs). Such an approach is simple, efficient, and adaptable to many sound recognition tasks. It has been superseded for accuracy in general-purpose sound recognition by more advanced methods (Stowell et al., 2015). We selected it to provide a common low-complexity baseline, and also because its simplicity meant it might successfully avoid overfitting to the training data that is avoid becoming overspecialised, given that the training and test data would have different characteristics.

The second baseline (code-named skfl) was a recent and more powerful classifier introduced for bird species recognition (Stowell & Plumbley, 2014a). This was the strongest audio-only bird species classifier in a 2014 international evaluation campaign. Relative to smacpy, it innovated in both the feature representation and the classification algorithm. The feature representation was an automatically learnt data transformation: two layers of "unsupervised feature learning" applied to mel spectrogram input—which is a spectrogram with its frequency axis warped to an approximation of human nonlinear frequency-band sensitivity. For classification, the method used a random forest, an ensemble learning method based on decision trees that has emerged as powerful and robust for many tasks in machine learning (Breiman, 2001). Both of these components are known to work well with difficult classification scenarios, such as multi-modal classes, unbalanced datasets, and outliers. We thus selected this second baseline as a representative of modern and flexible machine learning, designed for bird sounds. In principle, it could be more vulnerable to overfitting than the first baseline. However, because of the inherent difficulty of the task, we expected skfl to perform more strongly than smacpy, and to provide a high-performing baseline.

Python code for each of the baseline classifiers has previously been published (Stowell & Plumbley, 2014a; Stowell et al., 2015). Initial results using the baseline classifiers were published online as a guide to the challenge participants (http://machine-listening.eecs.qmul.ac.uk/2016/10/bird-audio-detection-baseline-generalisation/).

2.3 | The public challenge

Conduct of the challenge followed the design of previous successful contests on related topics: an open, public challenge in which teams were provided with standardised datasets, for developing and evaluating state-of-the-art methods against each other in a common framework (Goëau et al., 2016; Stowell et al., 2015). Having already determined the need for the challenge from remote monitoring literature and practitioners (Stowell, Wood, Stylianou, & Glotin, 2016), we announced intentions and led community discussion on the task design via a dedicated mailing list.

We collated our audio datasets into three portions: development data to be publicly shared (7,690 items of freefield1010 plus 8,000 items from Warblr), testing data whose true labels were to be kept private (10,000 items from Chernobyl plus 2,000 items from Warblr), and a separate set not used for the challenge itself but for further study of algorithm generalisation (PolandNFC). Providing two distinct development datasets allowed participants to test generalisation from one to the other, as part of their own algorithm development process, while keeping some datasets fully private allowed us to evaluate generalisation without concern that algorithms might have been configured to the specifics of a given dataset.

The public development datasets were distributed in September 2016, both audio and ground-truth annotations. Teams could then begin to develop methods and train their systems. In December 2016, we released the testing data (audio only), with a 1-month deadline for the submission of inferred detection labels. The short time horizon of 1 month was intended to minimise the opportunity for overly adapting the methods to the characteristics of the testing data. During this period, teams could make multiple submissions, but limited to a maximum of one per day. "Preview" results, calculated from 15% of the testing data, were provided in an interactive online plot, in order to give approximate feedback on performance (Figure S2).

Participants were allowed to run their software on their own machines and then to submit merely the outputs (as opposed to the software code), which our online system would then score without revealing the ground-truth labels for the testing data. Given that this open approach has potential vulnerabilities—such as recruiting manual labellers rather than developing automatic methods—we required the highest scoring teams to send in their code which we inspected and reran on our own systems, to ensure a fair outcome.

Apart from intrinsic motivation, incentives for participants were cash prizes: one for the strongest scoring system and one judges' award decided according to the use of interesting or novel methodology. This was done to stimulate conceptual development in the field, as opposed to the mere application of off-the-shelf deep learning. Participants were further required to submit technical notes describing their method, and later were invited to submit peer-reviewed conference papers to a special session at the European Signal Processing Conference (EUSIPCO) 2017. Of these, eight challenge-related papers were accepted and presented (Abrol et al., 2017; Adavanne, Drossos, Çakir, & Virtanen, 2017; Cakir, Adavanne, Parascandolo, Drossos, & Virtanen, 2017; Cakir, Parascandolo, Heittola, Huttunen, & Virtanen, 2017; Grill & Schlüter, 2017; Kong, Xu, & Plumbley, 2017; Pellegrini, 2017; Sandsten & Brynolfsson, 2017; Thakur, Jyothi, & Padmanabhan Rajan, 2017).

The challenge organisation was thus designed to achieve the following: public benchmarking of methods against a common task and data, specifically tailored to fully automatic configuration-free bird detection in unseen conditions; public documentation of the methods used to achieve leading results; and greater attention from machine learning researchers on data analysis tasks in environmental sound monitoring.

2.4 | Evaluation

Our goal was to evaluate algorithms for their ability to perform general-purpose bird detection, within the selected format of binary decisions for 10-s audio clips. A strong algorithm is one that can reliably separate the two classes "bird(s) present" and "no bird present." However, since our evaluation was general and not targeted at a specific application, we wished to generalise over the possible trade-offs of precision vs. recall (the relative cost of false positive detections vs. false negative detections). This strongly motivated our design such that participants should return probabilistic or graded outputs—a real-valued prediction for each audio clip rather than simply a 1 or 0—and our evaluation would use the well-studied AUC as the primary quality metric. The AUC measure has numerous qualities that make it well-suited to evaluation of such classification tasks: it generalises over all the possible thresholds that one might apply to real-valued detector outputs; unlike raw accuracy, it is not affected by "unbalanced" datasets having an uneven mixture of positive and negative items; chance performance for AUC is always 50% irrespective of dataset; and it has a probabilistic interpretation, as the probability that a given algorithm will rank a randomly selected positive instance more highly than a randomly selected negative instance (Fawcett, 2006).

The ranking interpretation just mentioned highlights another aspect of the AUC statistic: it treats detector outputs essentially as ranked values and is thus invariant to any monotonic mapping of the outputs, in particular to whether the outputs are well-calibrated probabilities or not. Well-calibrated, in this context, implies that when a detector outputs "0.75" for an item, this matches the empirical probability that in three out of four such cases the item is indeed a positive instance (Niculescu-Mizil & Caruana, 2005).

Thus, AUC does not evaluate calibration. But the need for calibration depends on the application: if a detector is being used to select a subset of strong detections, or to rank items for further manual inspection, there may be no need for calibration. However, if the detections are to be used in some probabilistic model, for example for modelling a population distribution, it is desirable for a detector to output well-calibrated probabilities. If a detector performs well in the sense evaluated by AUC, then its outputs can be mapped to probabilities by a post-processing step (Niculescu-Mizil & Caruana, 2005). Hence, we used AUC as our primary measure of quality, and separately we analysed the calibration of the submitted algorithms using the method of calibration plots, which are histogram plots comparing outputs against empirical probabilities (Niculescu-Mizil & Caruana, 2005). This approach does not modify the outputs of the classifiers—rather it analyses their predictions on the testing data, and the extent to which their probability values correspond with the empirical balance of positive and negative items.

To evaluate statistical significance and generate confidence intervals for the main outcomes, we used bootstrap sampling (Tibshirani & Efron, 1993; Urbano, 2013, Chapter 5). For this, we created 500 bootstrap resamples of the per-item predictions, from whose empirical statistics we calculated nonparametric significance tests (based on the distribution of AUC ranks across the 500 resamples) and confidence intervals (2.5 and 97.5 percentiles).

2.5 | Further analysis via PolandNFC dataset

After the challenge concluded, we took the highest scoring algorithm and applied it to the PolandNFC dataset, an unseen and difficult dataset containing night flight calls, often brief and distant. We used this in two

ways: (a) trained on a held-out portion (72.7%) of the PolandNFC data and tested on the remaining 27.3%; and (b) trained using the main challenge development data and again tested on the 27.3% of the PolandNFC data.

This allowed us to evaluate further the generalisation capability learnt by the network. For variant (a), the training dataset consisted of sixteen 30-min recordings collected over 11 nights (split into 28,784 1-s clips), and the testing dataset had six recordings from four nights (split into 10,793 1-s clips). Training and testing recording dates were disjoint sets. The testing set was held the same across variants (a) and (b) to ensure comparability of results. Important to note is the specific structure of the PolandNFC dataset—it contains mostly negatively annotated examples (only 3.2% for testing and 1.6% for training set were positive).

3 | RESULTS

In revalidating the testing set, we examined those items with the strongest mismatch between manual and automatic detection, to determine which was in error: 500 presumed negative and 1,243 presumed positive items. This showed inter-rater disagreement in 16.6% of such cases predominantly, the most ambiguous cases with barely audible bird sounds with amplitude close to the noise threshold. Note that this percentage is not representative of disagreement across the whole dataset, but only on the "controversial" cases. We also observed that a strong mismatch according to the automatic detectors did not necessarily imply human mislabelling: some perceptually obvious data items could be consistently misjudged by algorithms. We will discuss algorithm errors further below. Through revalidation, the inter-rater reliability, measured via the AUC, was measured as 96.7%. This value provides an approximate upper limit for machine performance since it reflects the extent of ambiguity in the data according to human listeners' perception.

The two baseline classifiers gave relatively good performance on the development data, the strongest at over 85% AUC in matched conditions, but generalised poorly. The simpler GMM-based baseline classifier showed consistently lower results than the more advanced classifier, as expected. It also showed strong resistance to overfitting in the sense that its performance on its training set was a very good predictor of its performance on a matched-conditions testing set. However, this was not sufficient to allow it to generalise to mismatched conditions, in which its performance degraded dramatically (Figure 1). The more advanced baseline classifier also degraded when tested in mismatched conditions, though to a lesser extent, attaining 74.8% AUC in the main evaluation.

3.1 | Challenge outcomes

Thirty different teams submitted results to the challenge, from various countries and research disciplines, with many submitting multiple times during the 1-month challenge period (Figure S2). Around half of the teams also submitted system descriptions, of which the majority were based on deep learning methods, often convolutional neural networks (CNNs) (Figure S1). To preprocess the audio for use in deep learning, most teams used a spectrogram representation—often a mel spectrogram, the same features as used in the skfl baseline. Many teams also used data augmentation, meaning that they artificially increased the amount of training data by copying and modifying data items in small ways, such as adding noise or shifting the audio in time. These strategies are in line with other work using machine learning for bird sound (Goëau et al., 2016; Salamon & Bello, 2017; Salamon et al., 2017).

Most teams were able to achieve over 80% AUC, but none over 90%: the strongest score was 88.7% AUC, attained by team "bulbul" (Thomas Grill) on the final day of challenge submission (Figure 1). The team has given further details of their approach in a short conference publication (Grill & Schlüter, 2017) and with open-source code

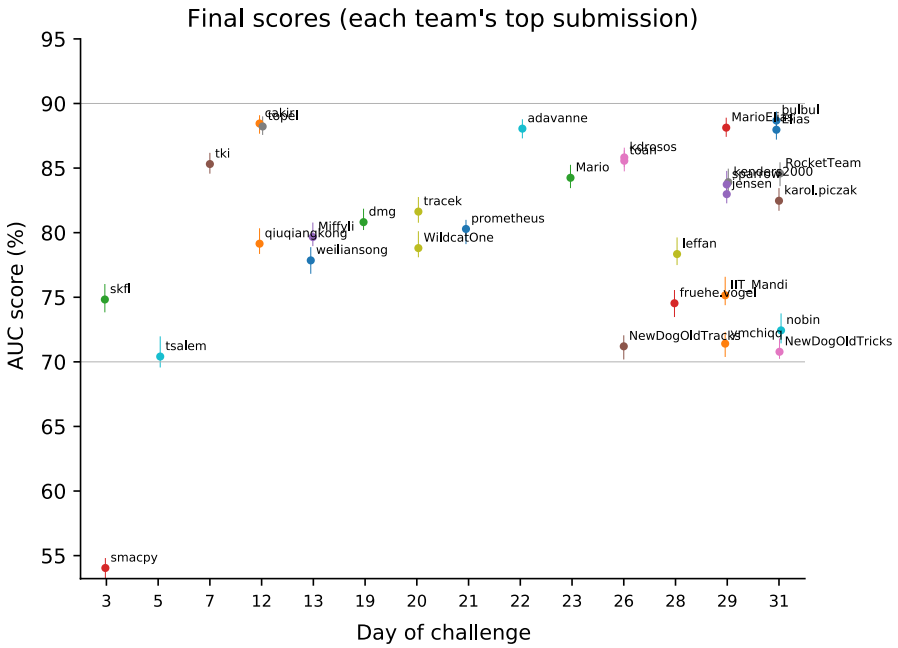

FIGURE 1 Final scores attained by the highest performing submission for each team. Error bars are estimated by bootstrap sampling. "skfl" and "smacpy" shown near the start are the baseline systems

available online (https://jobim.ofai.at/gitlab/gr/bird_audio_detection_challenge_2017/tree/master). A bootstrap test with a threshold of $p > 0.05$ showed that the results were compatible with any of the highest three teams having the most strongly performing system (cf. Figure S3). These were "bulbul" (Grill & Schlüter, 2017); "cakir" (Cakir, Adavanne, et al., 2017, Cakir, Parascandolo, et al., 2017), and "topel" (Pellegrini, 2017).

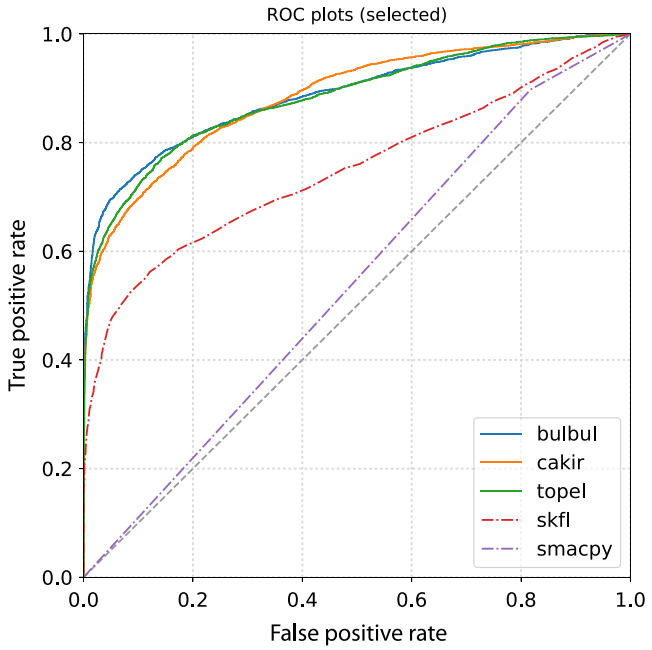

FIGURE 2 Receiver operating characteristic (ROC) plots for the systems attaining the three highest AUC scores and the baseline systems

AUC scores are summary statistics of ROC plots. ROC plots for systems were asymmetric (Figure 2), implying a spread of "difficulty" for the test items: there were some positive items that were easy to detect without incurring extra false positives as a side effect, while many remained difficult to detect. The most balanced of the ROC plots inspected was that of "cakir," implying a more even distribution of its discriminative power across the easy and difficult cases.

Since the testing data consisted of items from multiple "sites"—that is, known sites in the Chernobyl Exclusion Zone plus the Warblr (UK) data considered as a separate single site—we were able to calculate the AUC scores on a per-site basis (Figure 3, left). These showed a strong site dependency of algorithm performance. Whereas the Warblr data could be detected with an AUC of over 95%, Chernobyl sites showed varying difficulty for the detectors overall, some as low as 80% even for the leading algorithms. However, the overall AUC was very highly predictive of the average of the per-site AUCs (Pearson $R^2 = 0.80$; Figure 3, right). Note that the two AUC calculations are not independent and so some correlation is expected. The observed correlation validates that the overall AUC is usable as a summary of the per-site performance. The "bulbul" system remained the strongest performer even under the per-site analysis. The rank ordering of systems was not highly preserved: for example, the second-placed "cakir" system would have been ranked tenth if using the average per-site AUC.

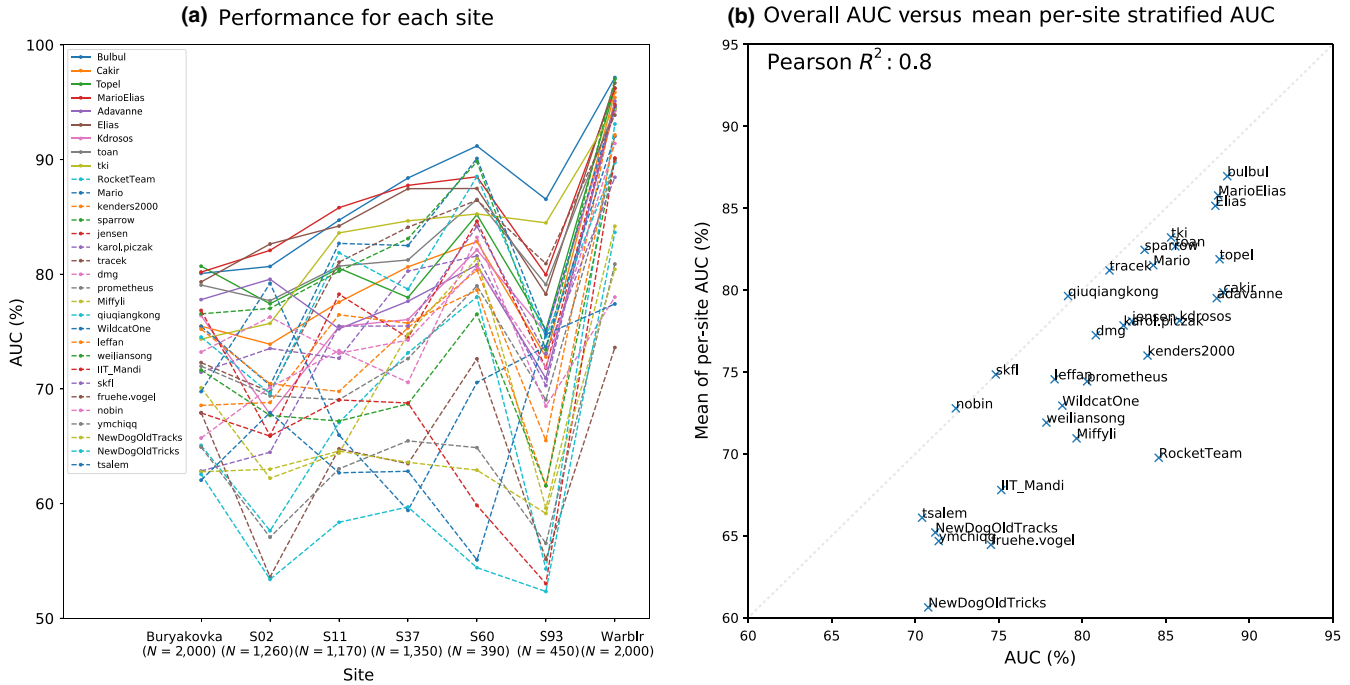

FIGURE 3 Performance (area under curve, AUC) of submitted algorithms analysed on a per-site basis. Warblr data (from around the UK) is treated as one site, while the other sites are recorders in the CEZ (Table 1). Left-hand plot shows the per-site results; right-hand plot shows how the AUC scores compare when calculated over the whole pooled dataset, vs. a mean of the stratified per-site AUCs

To visualise similarities and differences between system outputs, we applied PCA to the predictions that they produced. As an input, we used the predictions from each of the 30 systems treated as 12,000-dimensional vectors, one dimension per audio clip. A linear projection into two dimensions typically cannot represent all the variance in such high-dimensional data, but the linear constraint was found necessary for a non-trivial dimension reduction in this case with only 30 items. In the "similarity space" thus created, some of the lowest scoring systems formed two outlier clumps in terms of their predictions, while the stronger systems formed something of a continuum (Figure 4). The very strongest scoring systems did not cluster tightly together, indicating that there remains some diversity in the strategies implicit in these high-performing detectors.

We measured calibration curves separately for the Warblr and Chernobyl testing data (Figure 5). Calibration was generally better for Warblr, as one might expect given the availability of Warblr training data. Notably, the highest scoring submission "bulbul" had by far the worst calibration on the Chernobyl data: around 80% of cases it assigned a prediction value of 0.25 were indeed positive (vs. around 30% on the Warblr data). The second highest scoring submission "cakir" exhibited quite different behaviour, remaining relatively well calibrated even when assessing the unseen Chernobyl data.

3.2 | Error analysis

We inspected the 500 data items for which the predictions of the strongest systems exhibited mismatch with the revalidated ground truth, to characterise typical errors made by even the strongest machine learning systems in bird audio detection (Table 2). Such inspection is heuristic, relying on perceptual judgement to estimate the causes of errors; however, repeated tendencies give us indications about the performance of the current state of the art.

For false negatives, by far the most common observation was that positive items contained very faint bird sound (e.g., distant), often needing multiple listens to be sure it was present. These faint sounds had a low signal-to-noise ratio (SNR) and were often also quite reverberated. A low SNR was also a factor in the third most common presumed cause, noise masking. This category included general broadband "pink" noise sources including wind and rivers. There were other more specific categories of sound that appeared to act as masker or distractor causing systems to overlook the bird sound: insect noise was common in the CEZ data, while human sounds such as speech, whistling, or TVs were present in the Warblr data. The second most common presumed cause of false negatives was, however, the presence of extremely short calls: often a single "chink" sound, which might perhaps be overlooked or confused with rain-drop sounds. Some sounds were presumed to be missed because they were unusual for the dataset (e.g., goose honking), although this was not seen as a major factor.

False positives occurred at a much lower rate in the top 500 most mismatched items. They appeared to be caused in roughly equal proportion by insects, human sounds, and rain sounds (individual drops or diffuse rainfall).

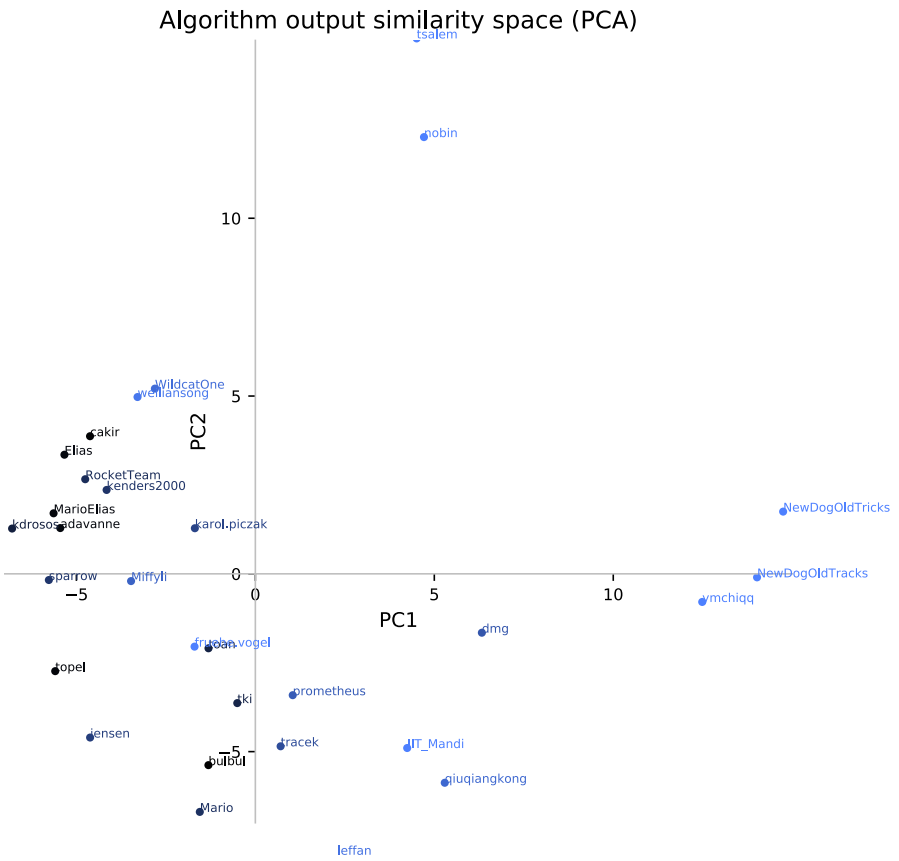

FIGURE 4 Similarity space comparing the top-scoring submission by each team (PCA projection of the submitted predictions after rank-transformation). Submissions are close together if their predictions were similar, irrespective of their accuracy. Submissions obtaining higher AUC scores are darker in colour. Variances explained: PC1 13%, PC2 9%

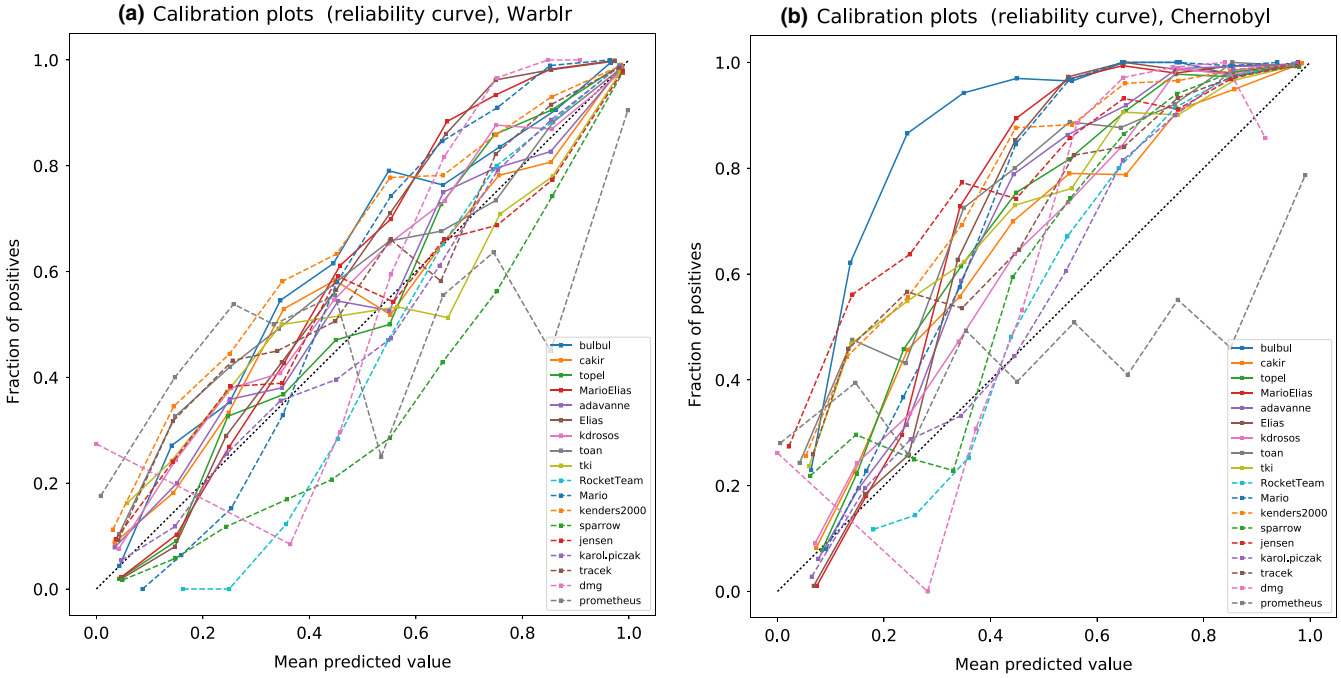

FIGURE 5 Calibration plots for the strongest submission by each team, separately for the Warblr test data (first plot) and Chernobyl test data (second plot). For legibility, we have limited this to the submissions attaining at least 80% AUC. A submission whose outputs are well-calibrated probabilities should yield a line close to the identity diagonal

TABLE 2 Inferred reasons for mistakes made by the strongest performing systems, annotated for the 500 items for which the systems showed the strongest deviation from ground truth. Note that the count data sum to more than 500, since multiple reasons could potentially be attributed to each item. "Clear" means the item was perceptually clear and should have been correctly labelled; "Dontknow" means that no obvious reason for a mistake is evident, even if the item is not particularly clear; all other rows are categories of presumed reasons for machine errors

Category	False positives	False negatives
Clear	1	68
Dontknow	7	40
Faint (e.g., v distant)	0	179
Short call	0	69
Noise-masking (including wind, river)	0	67
Insect	26	52
Human (speech, laughter, TV, imitation)	31	13
Rain (including drops)	26	5
Unusual bird sound	0	29
Miscellaneous distractor	0	11
Miscellaneous mammal	2	0

3.3 | Further analysis via PolandNFC dataset results

We then applied the highest scoring method ("bulbul") to the separate and unseen acoustic monitoring dataset PolandNFC. The bulbul algorithm consisted of two stages of inference: the first stage applied the pretrained neural network to make initial predictions; and the second stage then allowed the neural network to adapt to the observed data conditions, by feeding back the most confident predictions as new training data (Grill & Schlüter, 2017). We evaluated the outputs from each stage. AUC results were of high quality and were most strongly affected by the choice of training data: the matched-conditions training yielding much more accurate predictions (AUC = 95.0%) than the training performed on the bird audio detection Challenge dataset (AUC = 83.9%). The second stage adaptation offered AUC improvement to 87.8% in the case where the training set came from mismatched conditions. However, in matched-conditions training, the second stage actually incurred a slight reduction in performance, attaining 93.8% AUC. The detector also exhibited better calibration when trained in matched conditions, and the second stage retraining did not have a strong effect on calibration (Figure 6).

4 | DISCUSSION

Two broad observations emerge from this study:

1. Machine learning methods, primarily deep learning, are able to achieve very high recognition rates on remote monitoring acoustic data, despite weather noise, low SNRs, wide variation in bird call types, and even with mismatched training data. The AUC results presented here are a dramatic advance in the state of the art, and machine learning methods are of practical use in remote monitoring projects.

2. However, there remains a significant gap between performance in matched conditions and in mismatched conditions. True generalisation remains difficult and further work is needed. Projects are thus recommended to obtain some amount of matched-conditions training data where possible, and to treat automatic detection results with some caution especially with regard to the calibration of the outputs if they are to be treated as probabilities or if a fixed detection threshold is used. (Post-processing, such as Platt scaling, can ameliorate calibration issues Niculescu-Mizil & Caruana, 2005.) If a ranked-results approach is used (e.g., keeping the strongest *N* detections), which circumvents questions of calibration, then performance can remain strong even in mismatched conditions.

In practical applications, there are differing trade-offs of precision vs. recall of false negatives vs. false positives. The AUC statistics summarise over these. However, whatever trade-off is chosen, the current improvement in the state of the art provides dramatically reduced error rates (Figure 2), which corresponds, for example to a much lower amount of manual post-processing time in filtering out false positive results (Pamuła et al., 2017).

The strongest machine learning methods in this study were convolutional and/or recurrent neural nets (CNNs, RNNs, or CRNNs), as has been observed in other domains (LeCun et al., 2015). Perhaps the closest related domain is bird species classification from sound (Goëau et al., 2016; Knight et al., 2017; Salamon et al., 2017; Stowell & Plumbley, 2014a). Similar to the outcomes reported here, state-of-the-art methods in species classification use CNNs or CRNNs applied to spectrogram-type input data, as well as data augmentation to improve training, and some preprocessing of input spectrograms such as filtering or thresholding. The neural network architectures have much in common with those reported for our task. In fact, it is possible to design systems that attempt to address both detection and classification as two outputs from the same network (Morfi & Stowell, 2018). As the topic of deep learning for wildlife audio recognition continues to mature, we expect improved techniques to be applicable across all these related tasks.

Given our evaluation under conditions different from those in the training data, various participants explored self-adaptation, in which a trained network is fine-tuned upon exposure to the new conditions (without needing any additional ground-truth information) (Cakir, Adavanne, et al., 2017; Cakir, Parascandolo, et al., 2017; Grill & Schlüter, 2017). Participants reported mixed results of this, some observing no benefit. We found little benefit of self-adaptation for matched conditions; however, in cases where matched-conditions training data is not available, we found that it can reduce the adverse effect of the mismatch.

A further practical question is the feasibility of implementation on low-power devices for long-term deployment in the field. Deep learning experiments often require hardware acceleration, primarily for the training phase. After training, deep learning algorithms can be deployed onto smaller embedded units (Mac Aodha et al., 2018). However, the self-adaptation methods considered here are essentially additional rounds of training, albeit conducted with unlabelled data, and thus would incur quite some cost for use in the field. A pragmatic version of this would be to perform training or "pretraining" using mismatched data, then collecting a small amount of matched data from the target field conditions to perform self-adaptation, before fixing the algorithm parameters for use on a device.

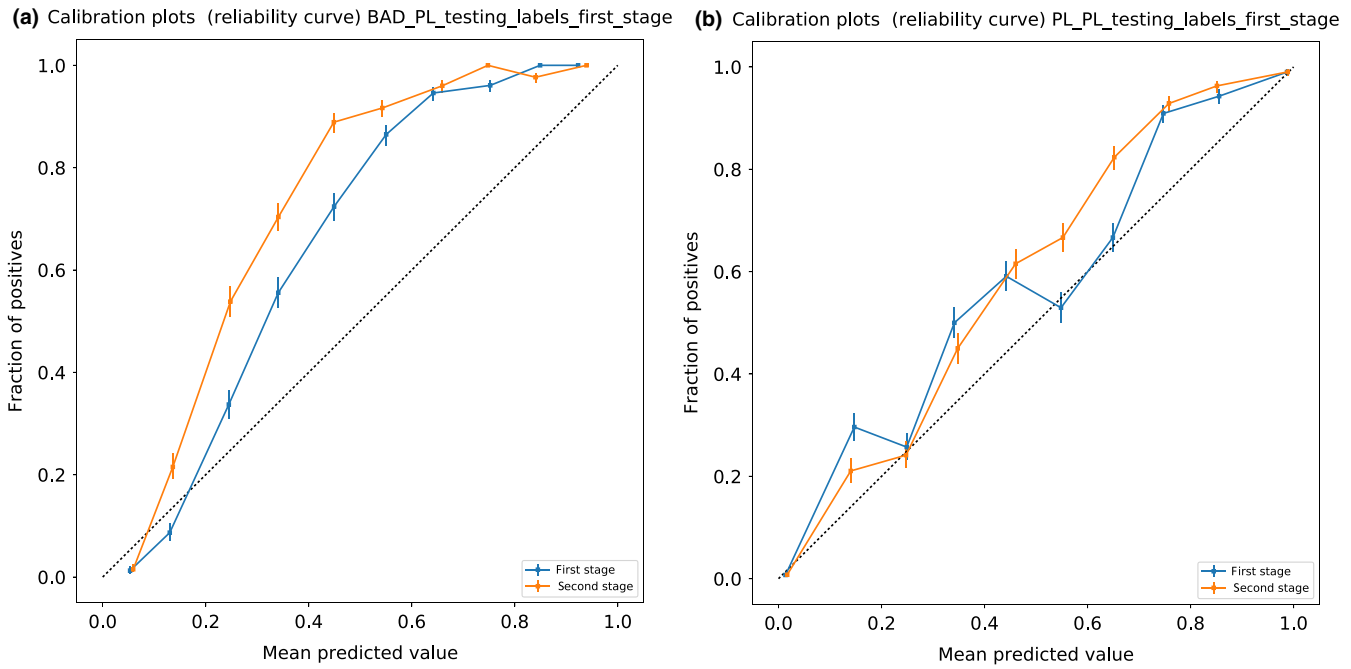

FIGURE 6 Calibration plots obtained when applying the highest scoring challenge system to remote monitoring audio data from Poland (cf. Figure 5). First plot: Challenge (mismatched) training; second plot: PolandNFC (matched) training. In these plots, 95% confidence intervals are calculated via the binomial method of Witten & Frank (2005, section 5.2.)

There remains a gap between human and machine performance. Our error analysis attributes this to sounds which present only weak evidence to the detector: often they were faint, reverberated, noise-masked, or very brief. Many of these were very hard to decide, as a listener. This perceptual difficulty, reflected in inter-rater disagreements, reminds us that some cases may be inherently ambiguous and thus may always be difficult for machine recognition. However, more than 1 in 10 of the top 500 items inspected were judged to be perceptually clear, meaning that the reason for those false negatives is due to a detector failing to model bird sound correctly, providing scope for algorithm improvements.

A further specific cause of detector errors stems from the ambiguity between very short "chink" bird calls and sounds such as individual rain drops which have similar effects in a spectrogram. A related issue was observed in bats, with the very short calls of species in the *Myotis* genus being the most difficult to disambiguate according to Walters et al. (2012). If the observable attributes of multiple sources overlap entirely, then it is not possible to distinguish them even in principle. However, at least in our case, human listeners can tell the difference, whether from context or from fine detail of signals.

How can we improve automatic detection on these weak sounds? Applying source separation and noise reduction is often unhelpful since weak sounds can be eliminated or distorted. More training data may be one answer; however, though our dataset sizes are smaller than those used in industrially backed application domains, we posit that larger training datasets alone would not close this gap. Instead, we expect further development of automatic pattern recognition to be a key. We look forward to algorithmic improvements such that detectors can use the full audio data as input (rather than spectrograms, which discard much information about temporal fine structure); can incorporate domain knowledge about the signals to be detected; and can make use of knowledge gained from other audio discrimination tasks (transfer learning or multitask learning).

Hutto and Stutzman (2009) previously performed an analysis of human sound detection of birds. Their comparison was between humans and "autonomous recording units"; however, note that in the latter case the detection was performed manually by inspecting spectrograms and listening to recordings, contrasted against a human listener in the field. Their results are thus not directly comparable to ours; however, they too found that distant bird sounds were the predominant cause of missed detections for remote sensing units. Furnas and Callas (2015) likewise studied in-field vs. audio-based detection using manual annotation, with similar results. They noted that detection probability could vary according to situational factors such as elevation and tree canopy cover. Digby, Towsey, Bell, and Teal (2013) evaluated automated detection in audio against in-field manual detection, for a single species (the little spotted kiwi *Apteryx owenii*), finding detection rates of around 40% with a relatively simple detection algorithm; despite this, they concluded that the high efficiency of automatic methods led to a large reduction in person-hours and thus was recommended. They found that wind noise exerted a larger influence on automatic detection than on manual detection.

Overall, our study design via a data challenge has been successful in moving forward the state of the art in acoustic remote monitoring. The design as a binary classification task, evaluated by AUC, is recommended as a way to generalise over some diversity in requirements among remote monitoring projects, with the calibration analysis as a useful addition to AUC evaluation. The use of multiple test sets sourced from different projects is a robust approach for general-purpose evaluation of algorithms, and we further recommend the use of per-site stratified AUC to account for per-site differences (The second edition of the Bird Audio Challenge, launched at time of writing, incorporates these recommendations, using per-site stratified AUC as well as adding further test sets to the challenge. http://dcase.community/challenge2018/task-bird-audio-detection). This complements the task-specific evaluation that a well-resourced individual project should undertake (cf. Knight et al., 2017). In some cases, going beyond the "yes/no" binary classification task is desirable to identify individual bird calls: the binary classification paradigm can in fact enable this, through a procedure of "weakly supervised learning" (Kong et al., 2017; Morfi & Stowell, 2017, 2018). In future evaluations, we recommend the exploration of such approaches, combining broad-scale detection with the elucidation of finer detail.

ACKNOWLEDGEMENTS

We thank Nick Beresford and Sergey Gashchak for their help with capturing the Chernobyl soundscape recordings, Paul Kendrick for preparation and annotation of Chernobyl data, Luciana Barçada for annotation of Chernobyl data, and Julien Ricard for programming and administering the challenge submission website. We also thank the many challenge participants for their enthusiastic effort. The Chernobyl data collection was undertaken within the TREE project (www.ceh.ac.uk/TREE), which is funded by the Natural Environment Research Council, the Environment Agency and Radioactive Waste Management Ltd.

AUTHORS' CONTRIBUTIONS

D.S., M.D.W., Y.S., and H.G. designed the data challenge study and its analysis; D.S., M.D.W., and H.P. provided datasets; H.G. led the creation of the challenge submission website; D.S. and H.P. performed tests of machine learning systems; D.S., H.G., and H.P. analysed the results; D.S. led the writing of the manuscript, with some sections by H.P. All authors contributed critically to the drafts and gave final approval for publication.

DATA ACCESSIBILITY

- Development datasets: both audio and annotations are available under CC-BY-4.0 licences.

warblrb10kaudio: https://archive.org/details/warblrb10k_public
ff1010bird audio: https://archive.org/details/ff1010bird
Annotations: https://doi.org/10.6084/m9.figshare.3851466.v1

- Testing data (Chernobyl/warblrb10k): audio public, but annotations held back for future challenges. Audio: https://archive.org/details/birdaudiodetectionchallenge_test
- PolandNFC data: held back, in preparation for future challenge.
- Source code for baseline classifiers:

GMM "smacpy" classifier (MIT licence): https://doi.org/10.5281/zenodo.1434195.

skfl feature-learning (MIT licence): https://doi.org/10.7717/peerj.488/supp-1.

- Source code for the online submission website (MIT licence): https://doi.org/10.5281/zenodo.1434268
- Source code and papers for various of the systems submitted by challenge teams are available via http://c4dm.eecs.qmul.ac.uk/events/badchallenge_results (various licences). The strongest submission "bulbul" is available at https://doi.org/10.5281/zenodo.1434624 and described further in Grill and Schlüter (2017).

REFERENCES

Abrol, V., Sharma, P., Thakur, A., Rajan, P., Dileep, A. D., & Sao, A. K. (2017). Archetypal analysis based sparse convex sequence kernel for bird activity detection. In *2017 25th European Signal Processing Conference (EUSIPCO)* (pp. 1774–1778). IEEE. https://doi.org/10.23919/eusipco.2017.8081514 Special Session on Bird Audio Signal Processing.

Adavanne, S., Drossos, K., Çakir, E., & Virtanen, T. (2017). Stacked convolutional and recurrent neural networks for bird audio detection. In *Signal Processing Conference (EUSIPCO), 2017 25th European* (pp. 1729–1733). IEEE. https://doi.org/10.23919/eusipco.2017.8081505 Special Session on Bird Audio Signal Processing.

Aide, T. M., Corrada-Bravo, C., Campos-Cerqueira, M., Milan, C., Vega, G., & Alvarez, R. (2013). Real-time bioacoustics monitoring and automated species identification. *PeerJ*, *1*, e103. https://doi.org/10.7717/peerj.103

Benetos, E., Stowell, D., & Plumbley, M. D. (2018). Approaches to complex sound scene analysis. In T. Virtanen, M. D. Plumbley, & D. Ellis (Eds.), *Computational analysis of sound scenes and events* (pp. 215–242). Cham: Springer. doi: 10. 1007/978-3-319-63450-0_8. https://doi.org/10.1007/978-3-319-63450-0

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A: 1010933404324

Cakir, E., Adavanne, S., Parascandolo, G., Drossos, K., & Virtanen, T. (2017). Convolutional recurrent neural networks for bird audio detection. In *Signal Processing Conference (EUSIPCO), 2017 25th European* (pp. 1744–1748). IEEE. https://doi.org/10.23919/eusipco.2017.8081508 Special Session on Bird Audio Signal Processing.

Cakir, E., Parascandolo, G., Heittola, T., Huttunen, H., & Virtanen, T. (2017). Convolutional recurrent neural networks for polyphonic sound event detection. *IEEE Transactions on Audio, Speech and Language Processing, Special Issue on Sound Scene and Event Analysis, arXiv preprint arXiv: 1702.06286*.

Colonna, J. G., Cristo, M., Júnior, M. S., & Nakamura, E. F. (2015). An incremental technique for real-time bioacoustic signal segmentation. *Expert Systems with Applications*, *42*, 7367–7374. https://doi.org/10.1016/j.eswa.2015.05.030

Digby, A., Towsey, M., Bell, B. D., & Teal, P. D. (2013). A practical comparison of manual and autonomous methods for acoustic monitoring. *Methods in Ecology and Evolution*, *4*, 675–683. https://doi.org/10.1111/2041-210X.12060

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*, 861–874. https://doi.org/10.1016/j.patrec. 2005.10.010

Frommolt, K.-H. (2017). Information obtained from long-term acoustic recordings: Applying bioacoustic techniques for monitoring wetland birds during breeding season. *Journal of Ornithology*, *158*, 1–10. https://doi.org/10.1007/s10336-016-1426-3

Furnas, B. J., & Callas, R. L. (2015). Using automated recorders and occupancy models to monitor common forest birds across a large geographic region. *The Journal of Wildlife Management*, *79*, 325–337. https://doi.org/10.1002/jwmg.821

Gashchak, S., Gulyaichenko, Y., Beresford, N. A., & Wood, M. D. (2017). European bison (*Bison bonasus*) in the Chornobyl Exclusion Zone (Ukraine) and prospects for its revival. *Proceedings of the Theriological School*, *15*, 58–66.

Goëau, H., Glotin, H., Vellinga, W.-P., Planque, R., & Joly, A. (2016). LifeCLEF bird identification task 2016: The arrival of deep learning. In *Working Notes of CLEF 2016-Conference and Labs of the Evaluation forum, Évora, Portugal*, 5-8 September (pp. 440–449).

Grill, T., & Schlüter, J. (2017). Two convolutional neural networks for bird detection in audio signals. In *Proceedings of EUSIPCO 2017* (pp. 1764–1768). https://doi.org/10.23919/eusipco.2017.8081512

Hill, A. P., Prince, P., Piña Covarrubias, E., Patrick Doncaster, C., Snaddon, J. L., & Rogers, A. (2017). AudioMoth: Evaluation of a smart open acoustic device for monitoring biodiversity and the environment. *Methods in Ecology and Evolution.*, *9*, 1199–1211. https://doi.org/10.1111%2F2041-210x.12955

Hutto, R. L., & Stutzman, R. J. (2009). Humans versus autonomous recording units: A comparison of point-count results. *Journal of Field Ornithology*, *80*, 387–398. https://doi.org/10.1111/j.1557-9263.2009.00245.x

Johnston, A., Ausden, M., Dodd, A. M., Bradbury, R. B., Chamberlain, D. E., Jiguet, F., … Rehfisch, M. M. (2013). Observed and predicted effects of climate change on species abundance in protected areas. *Nature Climate Change*, *3*, 1055–1061. https://doi.org/10.1038/NCLIMATE2035

Johnston, A., Newson, S. E., Risely, K., Musgrove, A. J., Massimino, D., Baillie, S. R., & Pearce-Higgins, J. W. (2014). Species traits explain variation in detectability of UK birds. *Bird Study*, *61*, 340–350. https://doi.org/10.1080/00063657.2014.941787

Joppa, L. N. (2017). Comment: The case for technology investments in the environment. *Nature*, *552*, 325–328. Retrieved from https://www.nature.com/articles/d41586-017-08675-7.

Kamp, J., Oppel, S., Heldbjerg, H., Nyegaard, T., & Donald, P. F. (2016). Unstructured citizen science data fail to detect long-term population declines of common birds in Denmark. *Diversity and Distributions*, *22*, 1024–1035. https://doi.org/10.1111/ddi.12463

Knight, E., Hannah, K., Foley, G., Scott, C., Brigham, R., & Bayne, E. (2017). Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs. *Avian Conservation and Ecology*, *12(2)*, 14. https://doi.org/10.5751/ACE-01114-120214

Kong, Q., Xu, Y., & Plumbley, M. D. (2017). Joint detection and classification convolutional neural network on weakly labelled bird audio detection. In *Signal Processing Conference (EUSIPCO), 2017 25th European* (pp. 1749–1753). IEEE. https://doi.org/10.23919/eusipco.2017.8081509 Special Session on Bird Audio Signal Processing.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444. https://doi.org/10.1038/nature14539

Mac Aodha, O., Gibb, R., Barlow, K. E., Browning, E., Firman, M., Freeman, R., … Pandourski, I. (2018). Bat detective-deep learning tools for bat acoustic signal detection. *PLoS Computational Biology*, *14*, 1–19. https://doi.org/10.1371/journal.pcbi.1005995

Marques, T. A., Thomas, L., Martin, S. W., Mellinger, D. K., Ward, J. A., Moretti, D. J., … Tyack, P. L. (2012). Estimating animal population density using passive acoustics. *Biological Reviews*, *88*, 287–309. https://doi.org/10.1111/brv.12001

Matsubayashi, S., Suzuki, R., Saito, F., Murate, T., Masuda, T., Yamamoto, K., & Okuno, H. G. (2017). Acoustic monitoring of the great reed

warbler using multiple microphone arrays and robot audition. *Journal of Robotics and Mechatronics*, *29*, 224–235. https://doi.org/10.20965/jrm.2017.p0224

Morfi, V., & Stowell, D. (2017) Deductive refinement of species labelling in weakly labelled birdsong recordings. In *Proc ICASSP 2017* (pp. 656–660). IEEE. https://doi.org/10.1109/icassp.2017.7952237

Morfi, V., & Stowell, D. (2018). Deep learning for audio transcription on low-resource datasets. *Applied Sciences*, *8(8)*, 1397. https://doi.org/10.3390/app8081397.

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning* (pp. 625–632). Bonn, Germany: ACM.

North American Bird Conservation Initiative. (2016). State of North America'sbirds2016.Technicalreport,Ottawa,Ontario.Retrievedfrom http://www.stateofthebirds.org/2016/state-of-the-birds-2016-pdf-download/.

Pamuła, H., Klaczynski, M., Remisiewicz, M., Wszolek, W., & Stowell, D. (2017). Adaptation of deep learning methods to nocturnal bird audio monitoring. In *LXIV Open Seminar on Acoustics (OSA) 2017*, Piekary Slaskie, Poland.

Pellegrini, T. (2017). Densely connected CNNs for bird audio detection. In *Proceedings of EUSIPCO 2017* (pp. 1734–1738). https://doi.org/10.23919/eusipco.2017.8081506 Special Session on Bird Audio Signal Processing.

RSPB. (2013). The state of nature in the UK and its overseas territories. Technical Report, RSPB and 24 other UK organisations. Retrieved from http://www.rspb.org.uk/ourwork/projects/details/363867-the-state-of-nature-report.

Salamon, J., & Bello, J. P. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, *24*, 279–283. https://doi.org/10.1109/LSP.2017.2657381

Salamon, J., Bello, J. P., Farnsworth, A., & Kelling, S. (2017). Fusing shallow and deep learning for bioacoustic bird species classification. In *Acoustics Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on* (pp. 141–145). IEEE. https://doi.org/10.1109/icassp.2017.7952134

Sandsten, M., & Brynolfsson, J. (2017). Classification of bird song syllables using Wigner-Ville ambiguity function cross-terms. In *Proceedings of EUSIPCO 2017* (pp. 1739–1743). https://doi.org/10.23919/eusipco.2017.8081507 Special Session on Bird Audio Signal Processing.

Stowell, D., Giannoulis, D., Benetos, E., Lagrange, M., & Plumbley, M. D. (2015). Detection and classification of acoustic scenes and events. *IEEE Transactions on Multimedia*, *17*, 1733–1746. https://doi.org/10.1109/TMM.2015. 2428998

Stowell, D., & Plumbley, M. D. (2014a). Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. *PeerJ*, *2*, e488. https://doi.org/10.7717/peerj.488

Stowell, D., & Plumbley, M. D. (2014b). An open dataset for research on audio field recording archives: Freefield1010. In *Proceedings of the Audio Engineering Society 53rd Conference on Semantic Audio (AES53)*. London, UK: Audio Engineering Society.

Stowell, D., Wood, M., Stylianou, Y., & Glotin, H. (2016). Bird detection in audio: A survey and a challenge. In *Proceedings of MLSP 2016*. https://doi.org/10.1109/mlsp.2016.7738875

Sugiyama, M., & Kawanabe, M. (2012). *Machine learning in non-stationary environments. Introduction to covariate shift adaptation*. Cambridge, MA: MIT Press.

Thakur, A., Jyothi, R., & Padmanabhan Rajan, A. D. (2017). Rapid bird activity detection using probabilistic sequence kernels. In *Proceedings of EUSIPCO 2017* (pp. 1754–1758). https://doi.org/10.23919/eusipco.2017.8081510 Special Session on Bird Audio Signal Processing.

Tibshirani, R. J., & Efron, B. (1993). An introduction to the bootstrap. *Monographs on Statistics and Applied Probability*, *57*, 1–436.

Towsey, M., Planitz, B., Nantes, A., Wimmer, J., & Roe, P. (2012). A toolbox for animal call recognition. *Bioacoustics*, *21*, 107–125. https://doi.org/10.1080/09524622.2011.648753

Urbano, J. (2013). *Evaluation in audio music similarity*. PhD Thesis, University Carlos III of Madrid. Retrieved from http://julian-urbano.info/publications/061-evaluation-audio-music-similarity

Walters, C. L., Freeman, R., Collen, A., Dietz, C., Fenton, M. B., Jones, G., … Parsons, S. (2012). A continental-scale tool for acoustic identification of European bats. *Journal of Applied Ecology*, *49*, 1064–1074. https://doi.org/10.1111/j.1365-2664.2012.02182.x

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques* (2nd ed.) San Francisco, CA: Morgan Kaufmann.

Wood, M., & Beresford, N. (2016). The wildlife of Chernobyl: 30 years without man. *Biologist*, *63*, 16–19.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Stowell D, Wood MD, Pamuła H, Stylianou Y, Glotin H. Automatic acoustic detection of birds through deep learning: The first Bird Audio Detection challenge. *Methods Ecol Evol*. 2018;00:1–13. https://doi.org/10.1111/2041-210X.13103